Carnegie Mellon University Language Technologies Institute

Better Synthetic Data by Retrieving and Transforming Existing Datasets

Saumya Gandhi*, Ritu Gala*, Vijay Viswanathan, Tongshuang Wu, and Graham Neubig

Problem

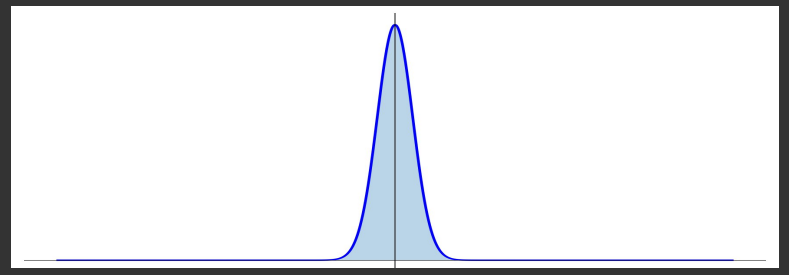
Generating synthetic data for new tasks is hard!

Problem

Generating synthetic data for new tasks is hard!

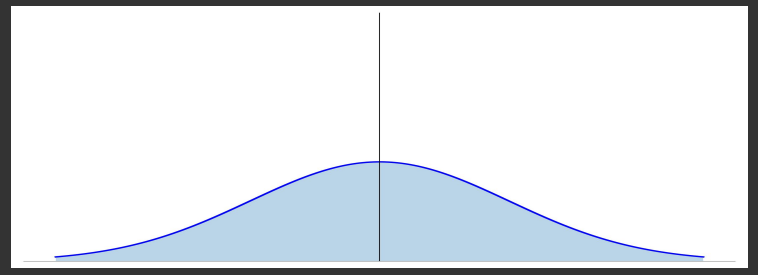
Diversity

Easy to generate a small amount of high-confidence examples



Quality

Easy to generate a *large amount* of *diverse* low-quality examples

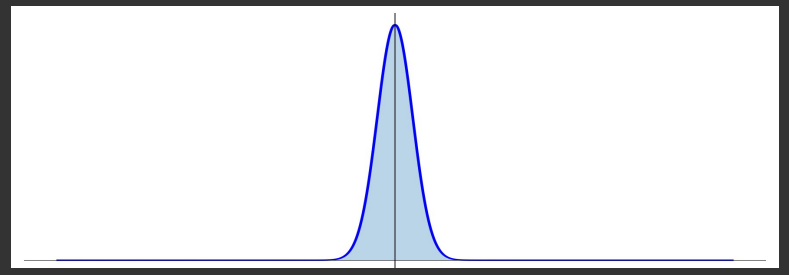


Problem

Generating synthetic data for new tasks is hard!

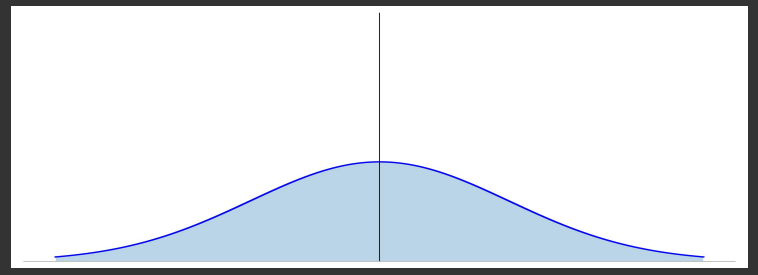
Diversity

Easy to generate a small amount of high-confidence examples



Quality

Easy to generate a *large amount* of *diverse* low-quality examples

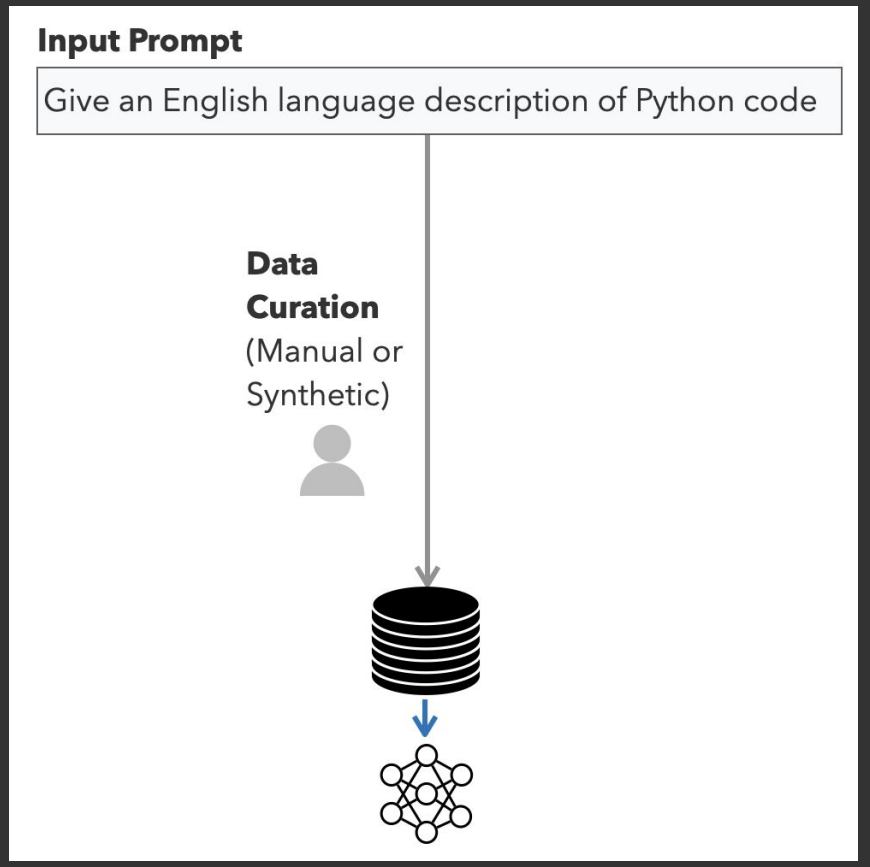


It's hard to achieve both!

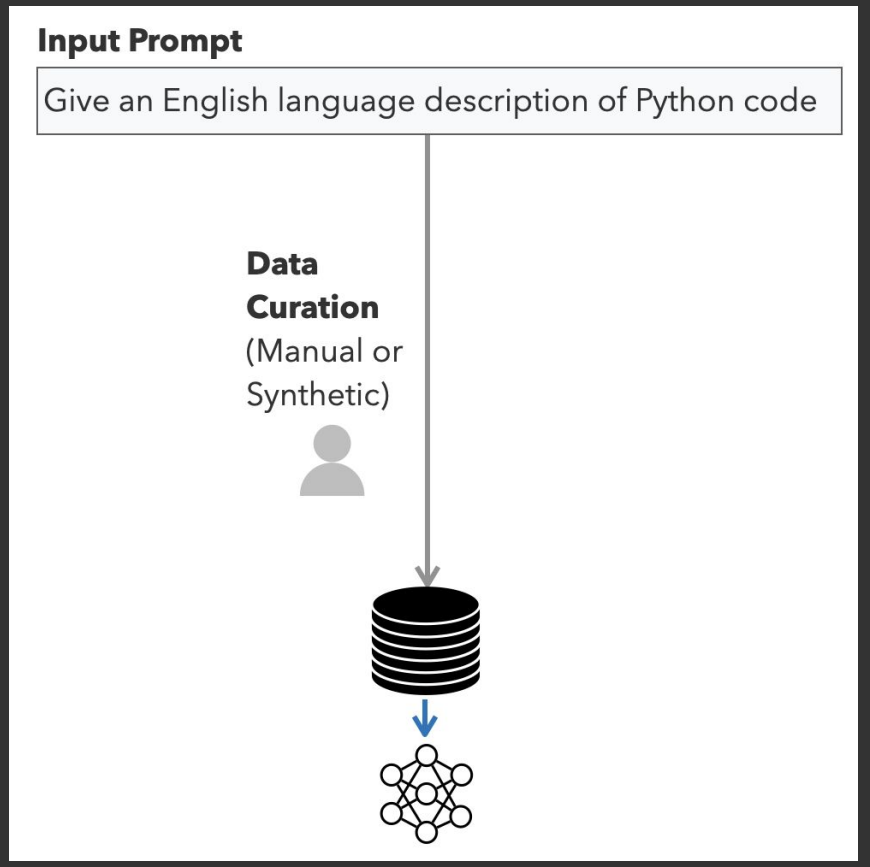
Main Idea

- Can we use *existing datasets* as a *starting point* for automatic synthetic data generation?

Traditional Approach to Synthetic Data Creation



Traditional Approach to Synthetic Data Creation



Leads to high cost / low diversity

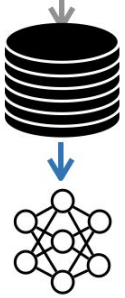
Traditional Approach to Synthetic Data Creation

Input Prompt

Give an English language description of Python code

Retrieval (cross-task transfer using existing datasets)

ID	Question	Solution	Test Cases	...
1	write an efficient algm for searching a number in sorted array	{binary search code}	in=[1,3,4,9],4 out=2	...
2	consider a circle...	{area_of_circle}	r=3, out=9	...
3	What is count of each char in a string	{str_char_count}	str="acaa" out={a:3, c:1}	...



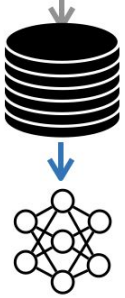
Traditional Approach to Synthetic Data Creation

Input Prompt

Give an English language description of Python code

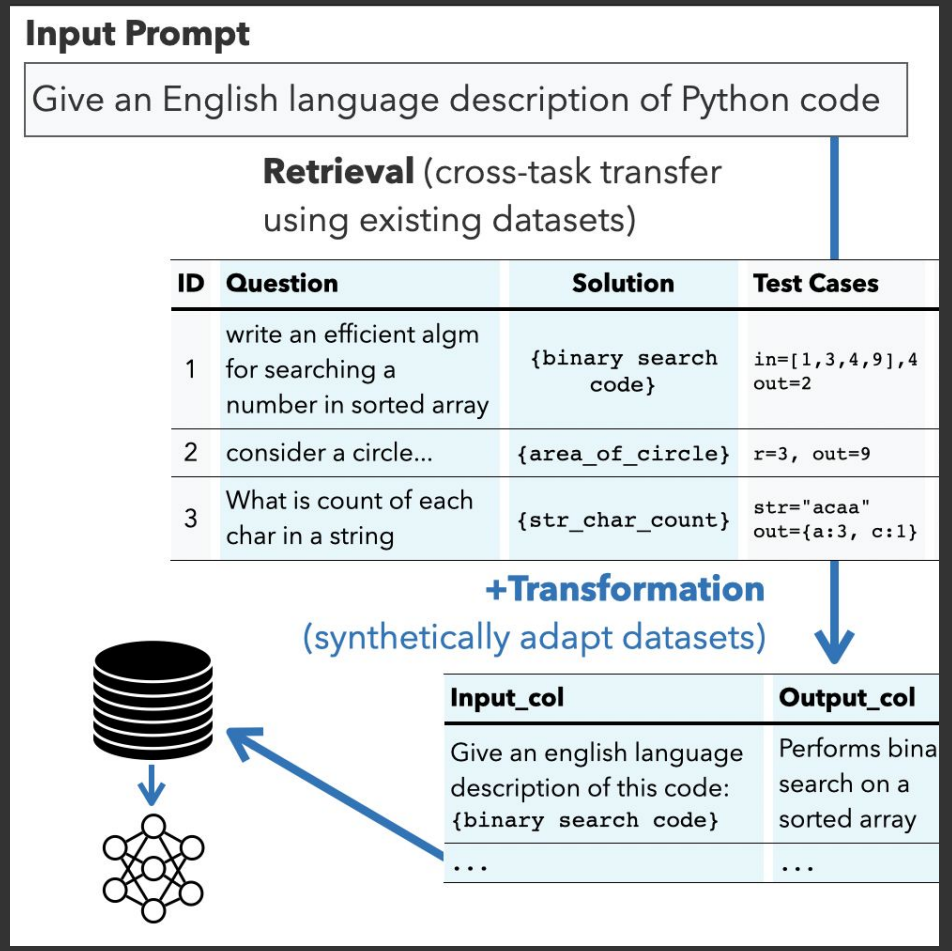
Retrieval (cross-task transfer using existing datasets)

ID	Question	Solution	Test Cases	...
1	write an efficient algm for searching a number in sorted array	{binary search code}	in=[1,3,4,9],4 out=2	...
2	consider a circle...	{area_of_circle}	r=3, out=9	...
3	What is count of each char in a string	{str_char_count}	str="acaa" out={a:3, c:1}	...

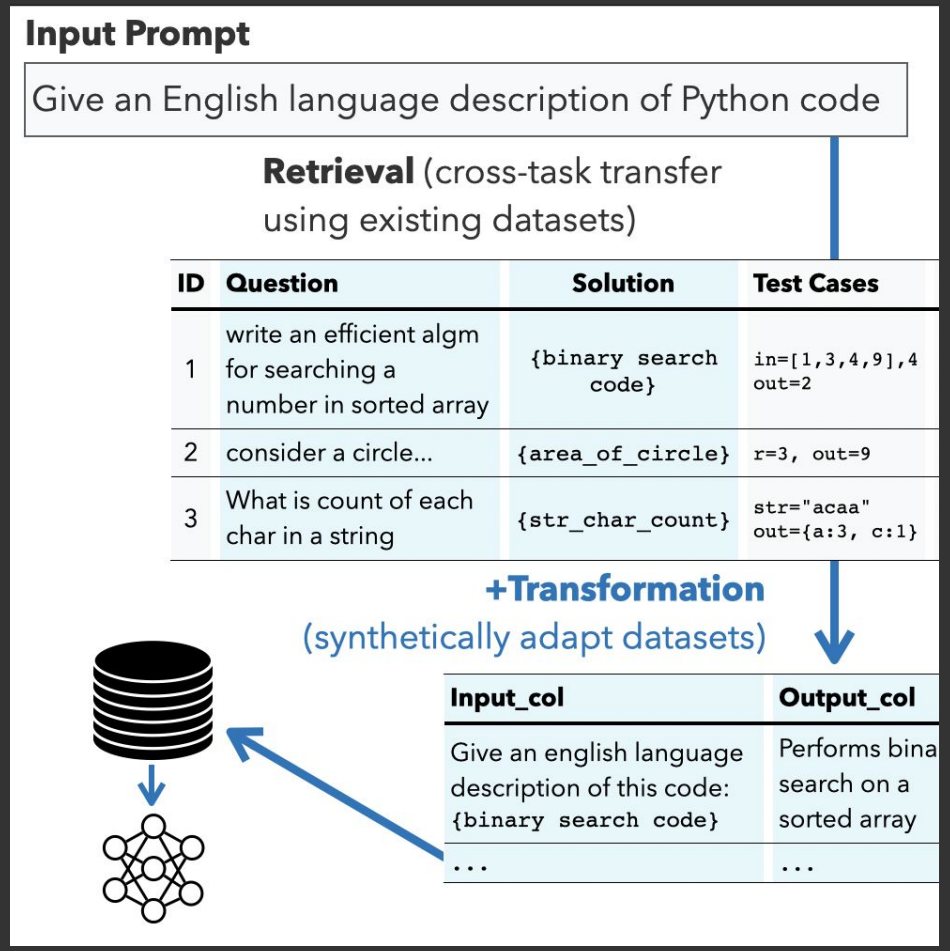


Need not follow the task exactly

DataTune: Synthetically transform retrieved datasets!

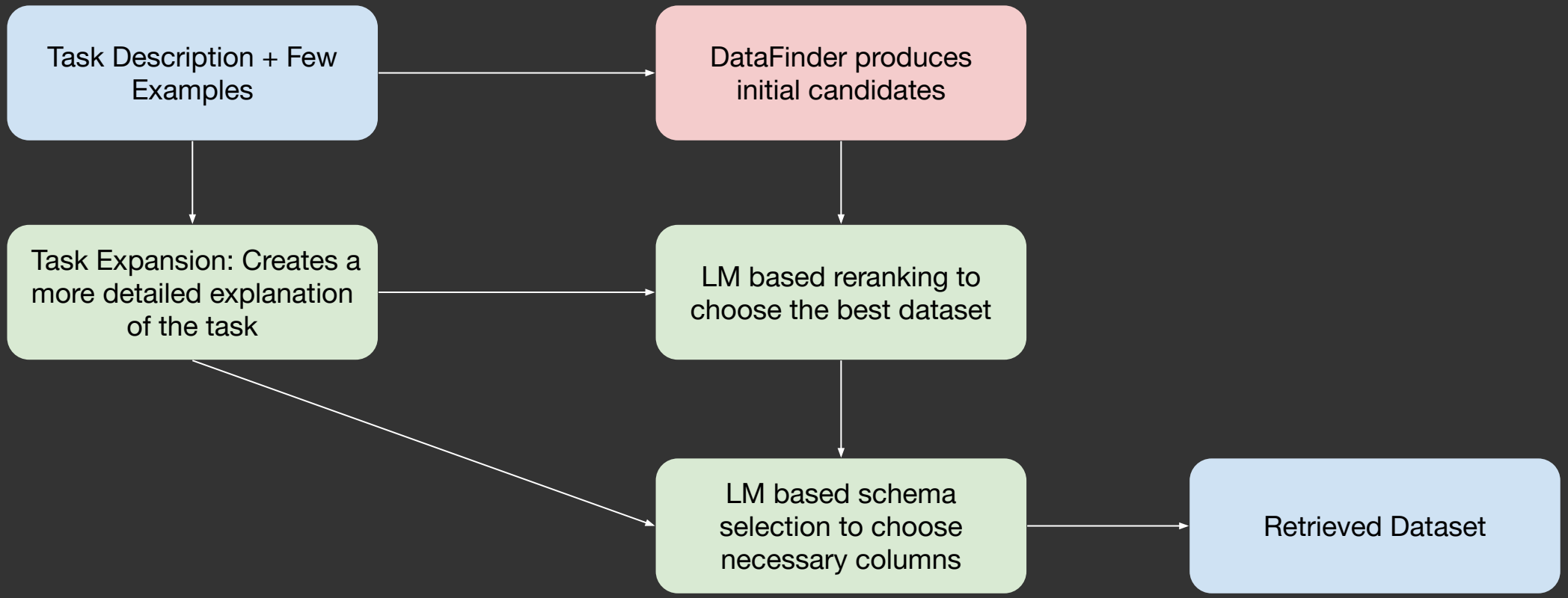


DataTune: Synthetically transform retrieved datasets!



Best of Both Worlds

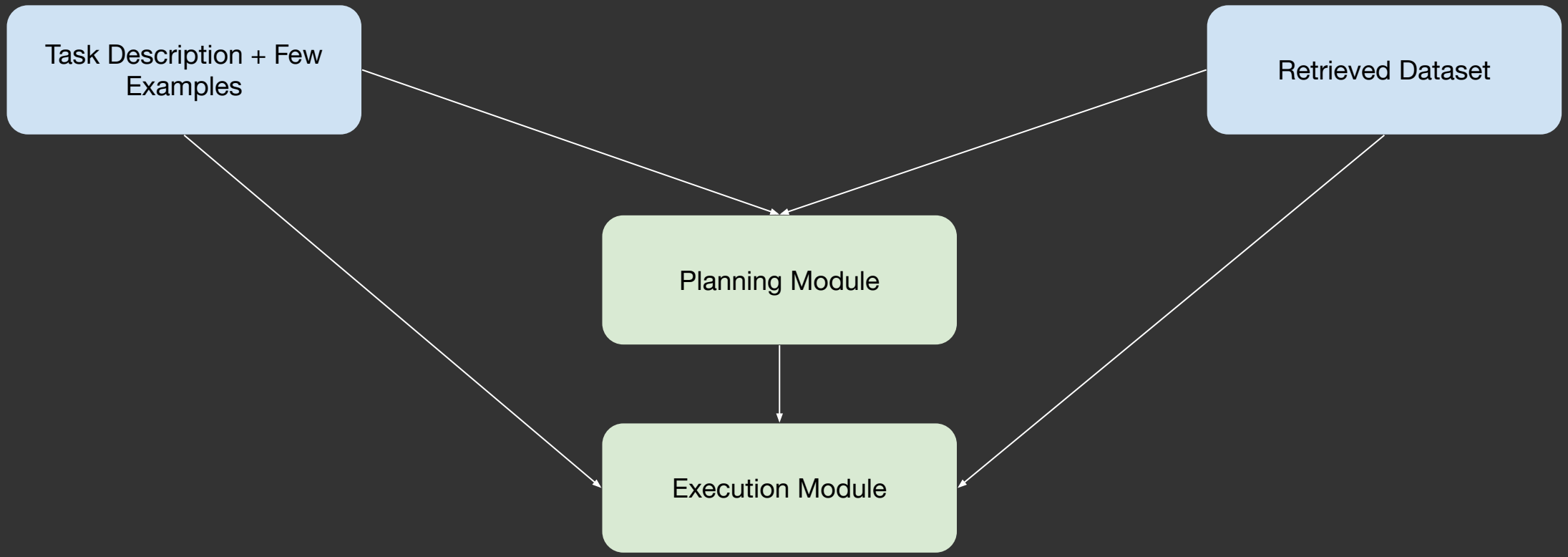
Dataset Retrieval



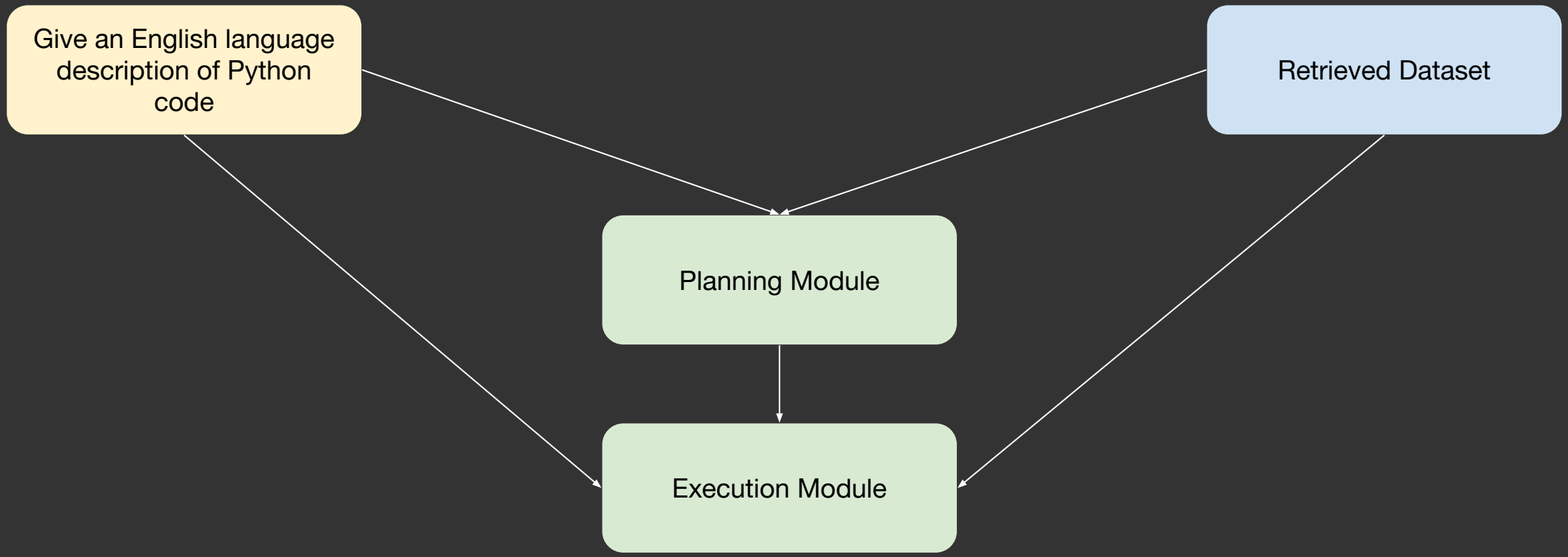
Vijay Viswanathan, Luyu Gao, Tongshuang Wu, Pengfei Liu, and Graham Neubig. 2023. **DataFinder: Scientific Dataset Recommendation from Natural Language Descriptions**. In *The 61st Annual Meeting of the Association for Computational Linguistics*.

https://aclanthology.org/2023.acl-long.573/

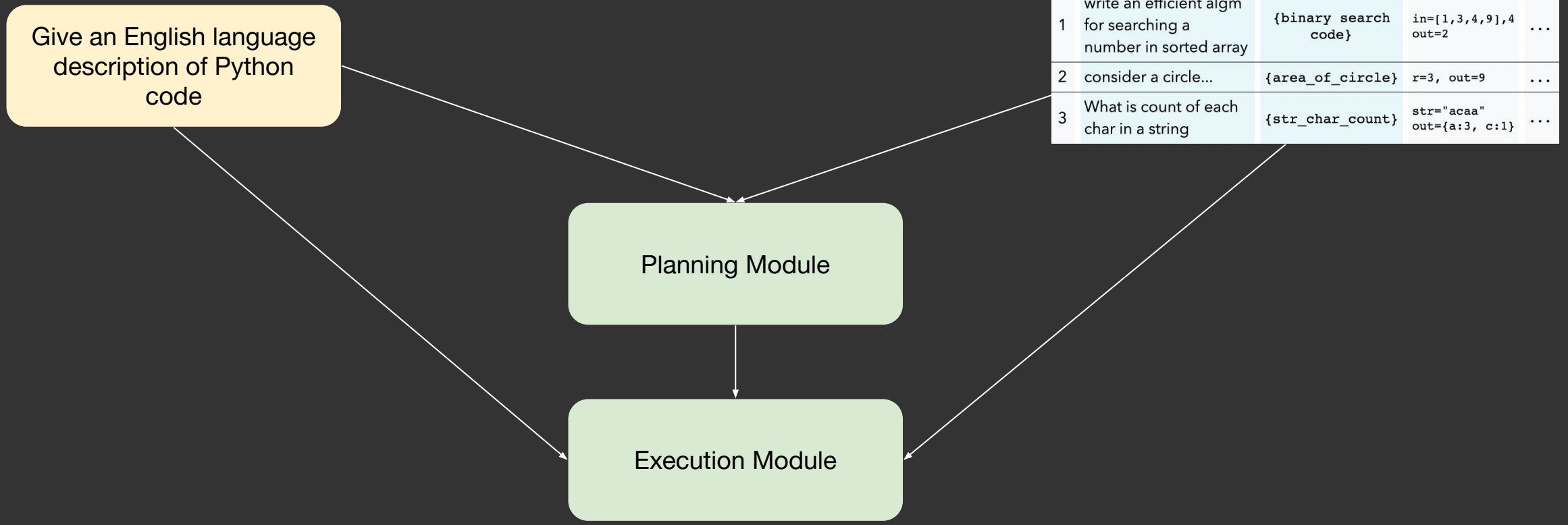
Dataset Transformation



Dataset Transformation Example



Dataset Transformation Example - Retrieved Dataset



ID	Question	Solution	Test Cases	...
1	write an efficient algm for searching a number in sorted array	{binary search code}	in=[1,3,4,9],4 out=2	...
2	consider a circle...	{area_of_circle}	r=3, out=9	...
3	What is count of each char in a string	{str_char_count}	str="acaa" out={a:3, c:1}	...

Dataset Transformation Example - Plan

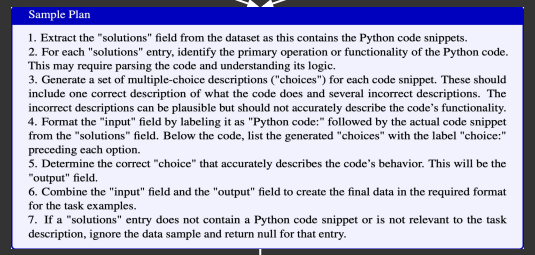
Sample Plan

1. Extract the "solutions" field from the dataset as this contains the Python code snippets.
2. For each "solutions" entry, identify the primary operation or functionality of the Python code. This may require parsing the code and understanding its logic.
3. Generate a set of multiple-choice descriptions ("choices") for each code snippet. These should include one correct description of what the code does and several incorrect descriptions. The incorrect descriptions can be plausible but should not accurately describe the code's functionality.
4. Format the "input" field by labeling it as "Python code:" followed by the actual code snippet from the "solutions" field. Below the code, list the generated "choices" with the label "choice:" preceding each option.
5. Determine the correct "choice" that accurately describes the code's behavior. This will be the "output" field.
6. Combine the "input" field and the "output" field to create the final data in the required format for the task examples.
7. If a "solutions" entry does not contain a Python code snippet or is not relevant to the task description, ignore the data sample and return null for that entry.

Dataset Transformation Example - Execution

Give an English language description of Python code

ID	Question	Solution	Test Cases	...
1	write an efficient algm for searching a number in sorted array	{binary search code}	in=[1,3,4,9],4 out=2	...
2	consider a circle...	{area_of_circle}	r=3, out=9	...
3	What is count of each char in a string	{str_char_count}	str="acaa" out={a:3, c:1}	...



Input_col	Output_col
Give an english language description of this code: {binary search code}	Performs binary search on a sorted array
...	...

Benchmark: BIG-Bench

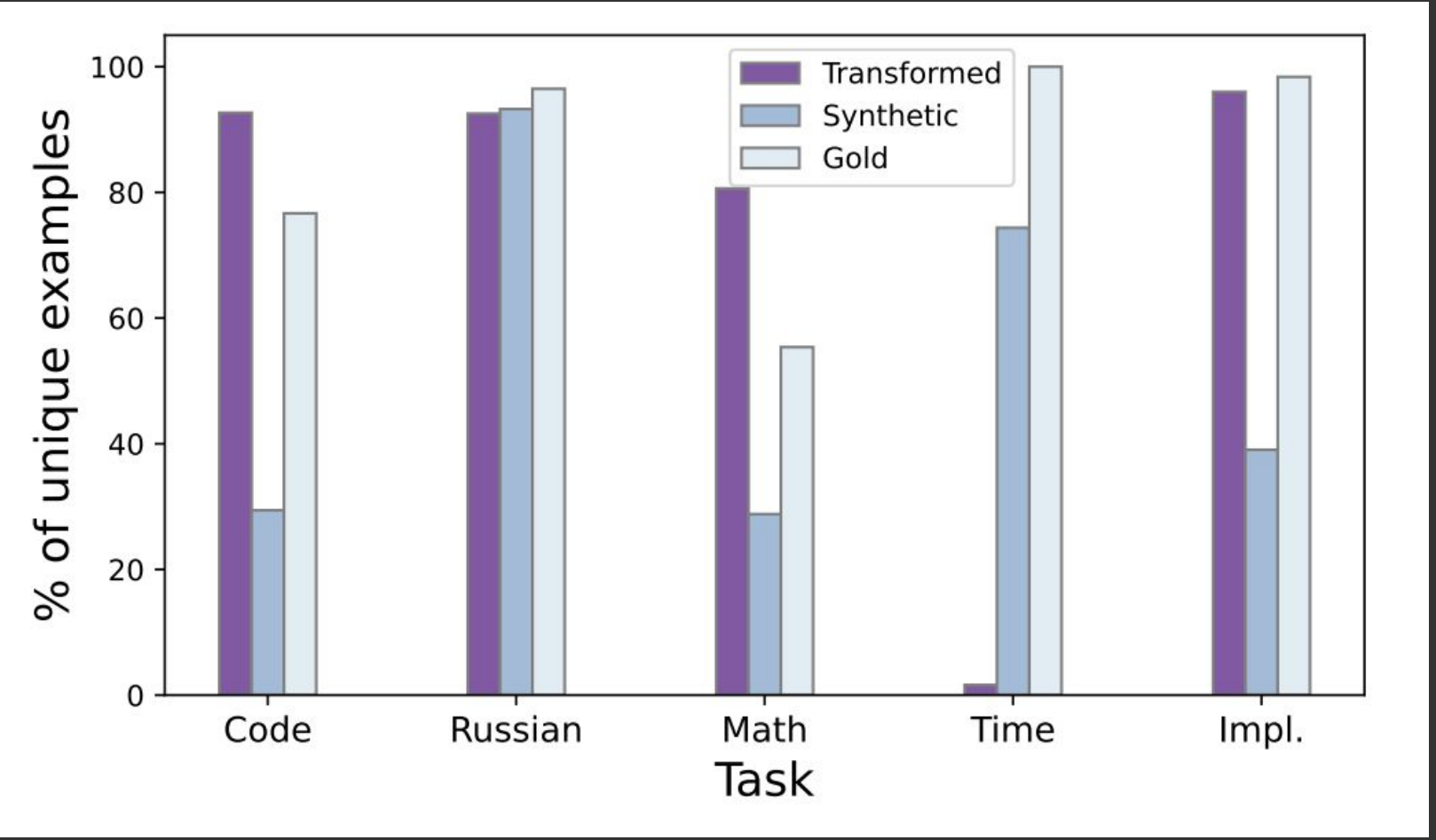
Task Name	Task Category	Abbreviation	Task Instruction
Temporal Sequences	Logical Reasoning	Time	Answer questions about which times certain events could have occurred.
Code Line Descriptions	Coding	Code	Give an English language description of Python code.
Elementary Math	Math	Math	Answer a multiple choice mathematical word problem.
Cause and Effect	Causal Reasoning	C&E	Answer multiple-choice questions distinguishing cause and effect.
Medical Questions in Russian	Domain Specific	Russian	Answer a yes/no question about medical text in Russian.
Implicatures	Contextual QA	Impl.	Predict whether Speaker 2's answer to Speaker 1 is affirmative or negative.

Results

	Method	Steps		Tasks						
		Retrieval Type	Generation	Time	Code	Math	C&E	Russian	Impl.	**Avg.**
	GPT-3.5 (few-shot)	-	-	50.6	75.6	30.4	96.7	90.6	64.2	68.0
	Mistral-7B (few-shot)	-	-	-2.5	62.3	2.9	37.2	39.8	39.0	29.8
Mistral-7B+	Existing data	Dense	-	-4.7	62.3	0.8	52.9	0.0	39.9	25.2
Mistral-7B+	Synthetic data	-	Synthetic	2.0	60.8	3.8	37.2	54.0	41.9	33.3
Mistral-7B+	DataTune	+ Reranker	Transformed	-2.1	71.2	1.3	**56.9**	48.0	41.9	36.2
Mistral-7B+	Prompt2Model	Dense	Synthetic	-2.0	73.4	4.7	33.8	**86.0**	44.0	40.0
Mistral-7B+	DataTune+Synthetic	+ Reranker	Both	**16.9**	**84.5**	**8.1**	41.2	68.0	**48.0**	**44.5**

- DataTune consistently outperforms few-shot prompting and other existing methods
- DataTune provides complementary benefits to purely synthetic generation

DataTune creates diverse datasets

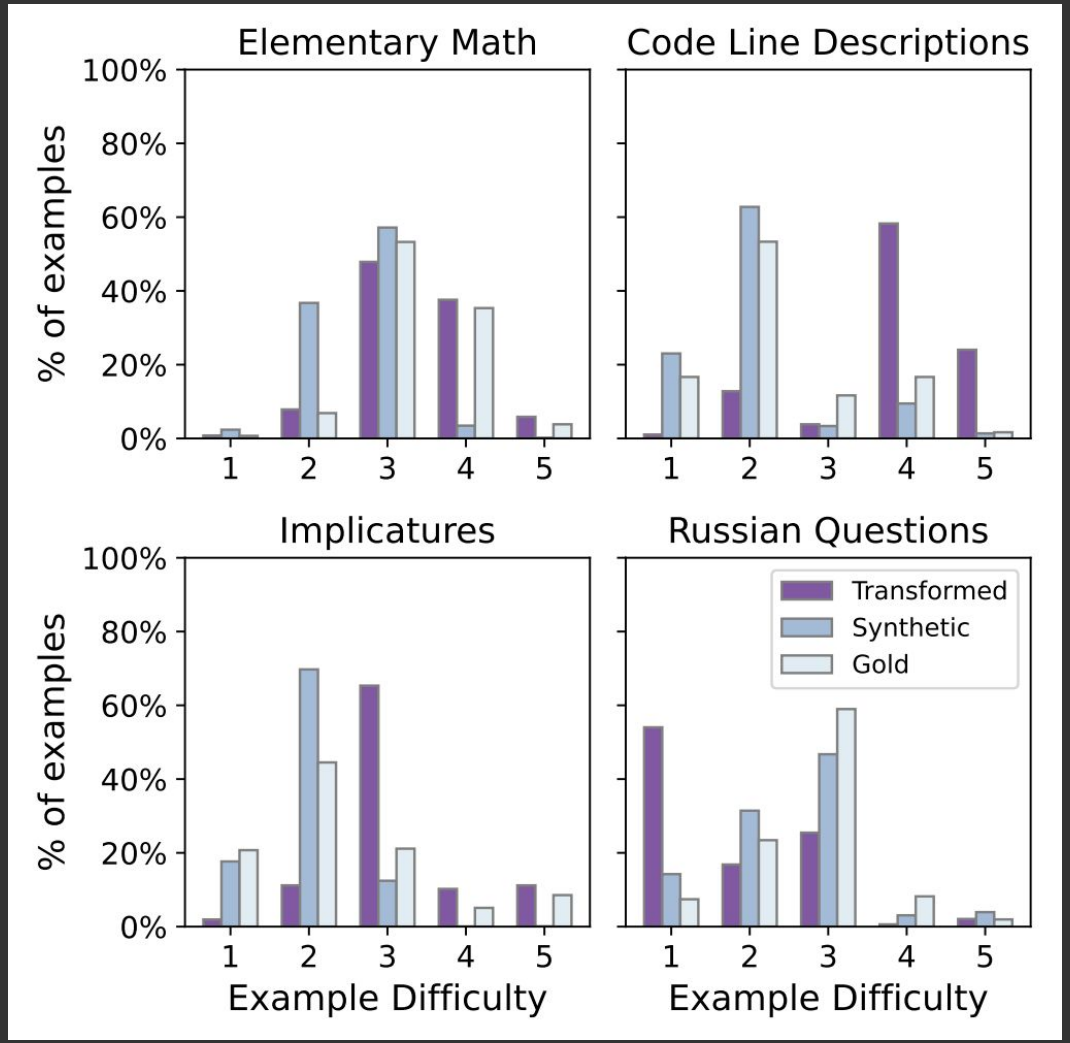


While still remaining correct

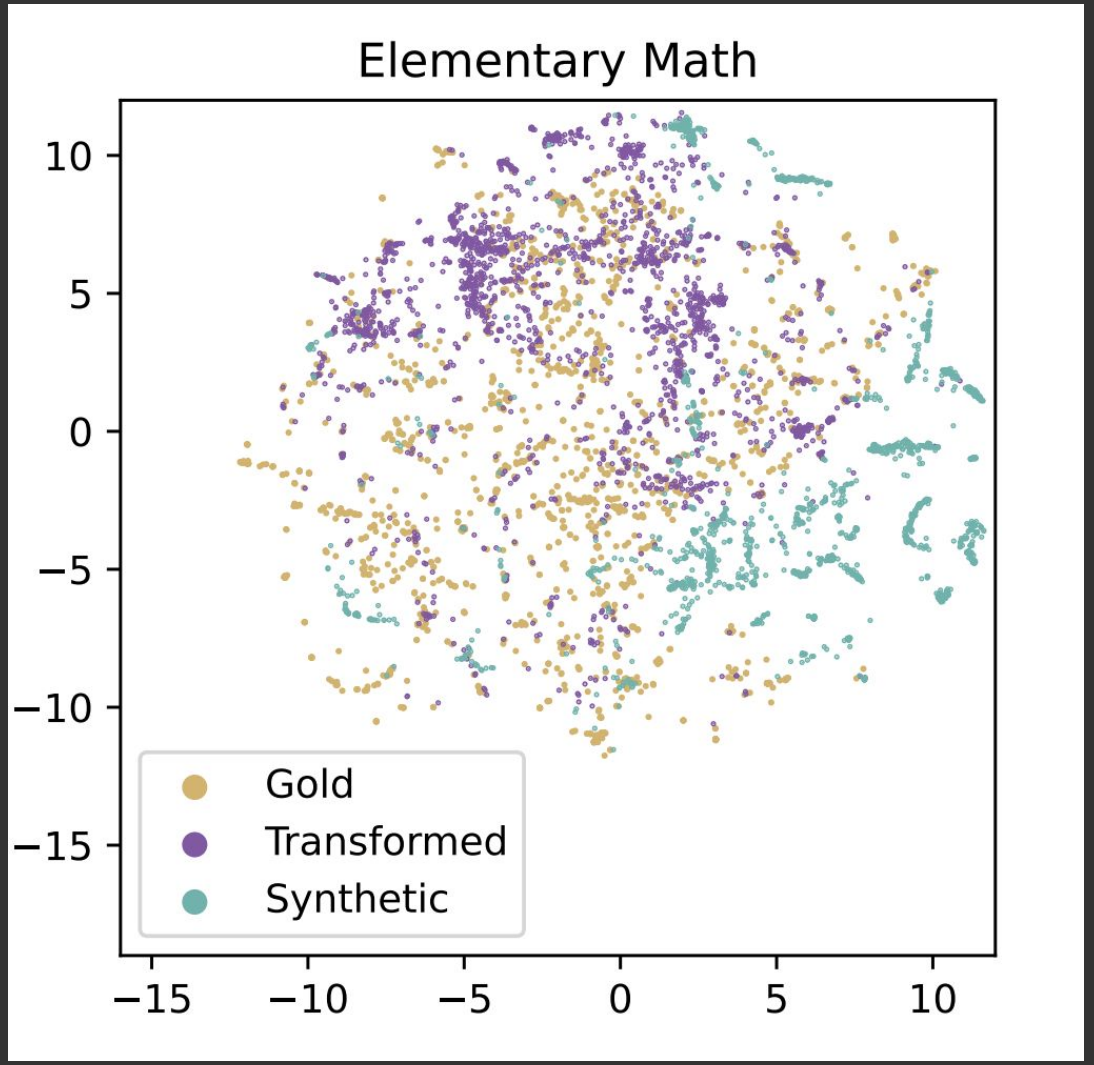
We performed human evaluation across 300 samples

- DataTune is correct 88% of the time
- Synthetic data creation is correct 86% of the time

And creating more difficult samples



Relation Between Synthetic and Transformed Datasets



Learn more about DataTune

Paper Link: https://arxiv.org/abs/2404.14361

Code Link: https://github.com/neulab/prompt2model